WIKIPEDIA

Bayesian information criterion

In statistics, the **Bayesian information criterion** (**BIC**) or **Schwarz information criterion** (also **SIC**, **SBC**, **SBIC**) is a criterion for model selection among a finite set of models; the model with the lowest BIC is preferred. It is based, in part, on the likelihood function and it is closely related to the Akaike information criterion (AIC).

When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in overfitting. Both BIC and AIC attempt to resolve this problem by introducing a penalty term for the number of parameters in the model; the penalty term is larger in BIC than in AIC.

The BIC was developed by Gideon E. Schwarz and published in a 1978 paper,[1] where he gave a Bayesian argument for adopting it.

Contents

Definition

The BIC is formally defined as[2][a]

$$\mathrm{BIC} = k\ln(n) - 2\ln(\widehat{L}).$$

where

- $\hat{L}$ = the maximized value of the likelihood function of the model M, i.e. $\hat{L} = p(x \mid \hat{\theta}, M)$, where $\hat{\theta}$ are the parameter values that maximize the likelihood function;
- x = the observed data;
- n = the number of data points in x, the number of observations, or equivalently, the sample size;

- k = the number of parameters estimated by the model. For example, in multiple linear regression, the estimated parameters are the intercept, the q slope parameters, and the constant variance of the errors; thus, $k = q + 2$.

Konishi and Kitagawa[4]:217 derive the BIC to approximate the distribution of the data, integrating out the parameters using Laplace's method, starting with the following model evidence:

$$p(x \mid M) = \int p(x \mid \theta, M)\pi(\theta \mid M)\, d\theta$$

where $\pi(\theta \mid M)$ is the prior for θ under model M.

The log-likelihood, $\ln(p(x|\theta, M))$, is then expanded to a second order Taylor series about the MLE, $\hat{\theta}$, assuming it is twice differentiable as follows:

$$\ln(p(x \mid \theta, M)) = \ln(\widehat{L}) - 0.5(\theta - \hat{\theta})' n\mathcal{I}(\theta)(\theta - \hat{\theta}) + R(x, \theta),$$

where $\mathcal{I}(\theta)$ is the average observed information per observation, and prime ($'$) denotes transpose of the vector $(\theta - \hat{\theta})$. To the extent that $R(x, \theta)$ is negligible and $\pi(\theta \mid M)$ is relatively linear near $\hat{\theta}$, we can integrate out θ to get the following:

$$p(x \mid M) \approx \hat{L}(2\pi/n)^{k/2}|\mathcal{I}(\hat{\theta})|^{-1/2}\pi(\hat{\theta})$$

As n increases, we can ignore $|\mathcal{I}(\hat{\theta})|$ and $\pi(\hat{\theta})$ as they are $O(1)$. Thus,

$$p(x \mid M) = \exp\{\ln \widehat{L} - (k/2)\ln(n) + O(1)\} = \exp(-\mathrm{BIC}/2 + O(1)),$$

where BIC is defined as above, and $\widehat{L}$ either (a) is the Bayesian posterior mode or (b) uses the MLE and the prior $\pi(\theta \mid M)$ has nonzero slope at the MLE. Then the posterior

$$p(M \mid x) \propto p(x \mid M)p(M) \approx \exp(-\mathrm{BIC}/2)p(M)$$

Properties

- It is independent of the prior.
- It can measure the efficiency of the parameterized model in terms of predicting the data.
- It penalizes the complexity of the model where complexity refers to the number of parameters in the model.
- It is approximately equal to the minimum description length criterion but with negative sign.
- It can be used to choose the number of clusters according to the intrinsic complexity present in a particular dataset.
- It is closely related to other penalized likelihood criteria such as Deviance information criterion and the Akaike information criterion.

Limitations

The BIC suffers from two main limitations[5]

1. the above approximation is only valid for sample size n much larger than the number k of parameters in the model.
2. the BIC cannot handle complex collections of models as in the variable selection (or feature selection) problem in high-dimension.[5]

Gaussian special case

Under the assumption that the model errors or disturbances are independent and identically distributed according to a normal distribution and that the boundary condition that the derivative of the log likelihood with respect to the true variance is zero, this becomes (*up to an additive constant*, which depends only on n and not on the model):[6]

$$\mathrm{BIC} = n\ln(\widehat{\sigma_e^2}) + k\ln(n)$$

where $\widehat{\sigma_e^2}$ is the error variance. The error variance in this case is defined as

$$\widehat{\sigma_e^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{x_i})^2.$$

which is a biased estimator for the true variance.

In terms of the residual sum of squares (RSS) the BIC is

$$\mathrm{BIC} = n\ln(RSS/n) + k\ln(n)$$

When testing multiple linear models against a saturated model, the BIC can be rewritten in terms of the deviance χ^2 as:[7]

$$\mathrm{BIC} = \chi^2 + k\ln(n)$$

where k is the number of model parameters in the test.

When picking from several models, the one with the lowest BIC is preferred. The BIC is an increasing function of the error variance σ_e^2 and an increasing function of k. That is, unexplained variation in the dependent variable and the number of explanatory variables increase the value of BIC. Hence, lower BIC implies either fewer explanatory variables, better fit, or both. The strength of the evidence against the model with the higher BIC value can be summarized as follows:[7]

ΔBIC	Evidence against higher BIC
0 to 2	Not worth more than a bare mention
2 to 6	Positive
6 to 10	Strong
>10	Very strong

The BIC generally penalizes free parameters more strongly than the Akaike information criterion, though it depends on the size of n and relative magnitude of n and k.

It is important to keep in mind that the BIC can be used to compare estimated models only when the numerical values of the dependent variable[b] are identical for all models being compared. The models being compared need not be nested, unlike the case when models are being compared using an F-test or a likelihood ratio test.

See also

- Akaike information criterion
- Bayes factor
- Bayesian model comparison
- Deviance information criterion
- Hannan–Quinn information criterion
- Jensen–Shannon divergence
- Kullback–Leibler divergence
- Minimum message length

Notes

a. The AIC, AICc and BIC defined by Claeskens and Hjort[3] are the negatives of those defined in this article and in most other standard references.
b. A dependent variable is also called a response variable or an outcome variable. See Regression analysis.

References

1. Schwarz, Gideon E. (1978), "Estimating the dimension of a model", Annals of Statistics, **6** (2): 461–464, doi:10.1214/aos/1176344136 (https://doi.org/10.1214%2Faos%2F1176344136), MR 0468014 (https://www.ams.org/mathscinet-getitem?mr=0468014).
2. Wit, Ernst; Edwin van den Heuvel; Jan-Willem Romeyn (2012). " 'All models are wrong...': an introduction to model uncertainty" (https://pure.rug.nl/ws/files/13270992/2012StatistNeerlWit.pdf) (PDF). Statistica Neerlandica. **66** (3): 217–236. doi:10.1111/j.1467-9574.2012.00530.x (https://doi.org/10.1111%2Fj.1467-9574.2012.00530.x).
3. Claeskens, G.; Hjort, N. L. (2008), Model Selection and Model Averaging, Cambridge University Press
4. Konishi, Sadanori; Kitagawa, Genshiro (2008). Information criteria and statistical modeling. Springer. ISBN 978-0-387-71886-6.
5. Giraud, C. (2015). Introduction to high-dimensional statistics. Chapman & Hall/CRC. ISBN 9781482237948.
6. Priestley, M.B. (1981). Spectral Analysis and Time Series. Academic Press. ISBN 978-0-12-564922-3. (p. 375).
7. Kass, Robert E.; Raftery, Adrian E. (1995), "Bayes Factors", Journal of the American Statistical Association, **90** (430): 773–795, doi:10.2307/2291091 (https://doi.org/10.2307%2F2291091), ISSN 0162-1459 (https://www.worldcat.org/issn/0162-1459), JSTOR 2291091 (https://www.jstor.org/stable/2291091).

Further reading

- Bhat, H. S.; Kumar, N (2010). "On the derivation of the Bayesian Information Criterion" (https://web.archive.org/web/20120328065032/http://nscs00.ucmerced.edu/~nkumar

4/BhatKumarBIC.pdf) (PDF). Archived from the original (http://nscs00.ucmerced.edu/~nkumar4/BhatKumarBIC.pdf) (PDF) on 28 March 2012.

- Findley, D. F. (1991). "Counterexamples to parsimony and BIC". Annals of the Institute of Statistical Mathematics. **43** (3): 505–514. doi:10.1007/BF00053369 (https://doi.org/10.1007%2FBF00053369).
- Kass, R. E.; Wasserman, L. (1995). "A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion". Journal of the American Statistical Association. **90** (431): 928–934. doi:10.2307/2291327 (https://doi.org/10.2307%2F2291327). JSTOR 2291327 (https://www.jstor.org/stable/2291327).
- Liddle, A. R. (2007). "Information criteria for astrophysical model selection". Monthly Notices of the Royal Astronomical Society. **377** (1): L74–L78. arXiv:astro-ph/0701113 (https://arxiv.org/abs/astro-ph/0701113). Bibcode:2007MNRAS.377L..74L (https://ui.adsabs.harvard.edu/abs/2007MNRAS.377L..74L). doi:10.1111/j.1745-3933.2007.00306.x (https://doi.org/10.1111%2Fj.1745-3933.2007.00306.x).
- McQuarrie, A. D. R.; Tsai, C.-L. (1998). Regression and Time Series Model Selection. World Scientific.

External links

- Information Criteria and Model Selection (http://personal.psu.edu/hxb11/INFORMATIONCRIT.PDF)
- Sparse Vector Autoregressive Modeling (https://arxiv.org/abs/1207.0520)